Characterising AI Weakness in Detecting Personal Data from Images By Crowds

Master Thesis

Ashay Somai

Delft University of Technology

TU Delft

Characterising AI Weakness in Detecting Personal Data from Images By Crowds

Master Thesis

by

Ashay Somai

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday December 21, 2021 at 9:00 AM.

Student number:	4366220	
Project duration:	November 20, 2020 – December 13, 2021	
Thesis committee:	Prof. dr. ir. Geert-Jan Houben,	Web Information Systems, chair
	Dr. ir. Jie Yang,	Web Information Systems, daily supervisor
	Dr. Qing Wang,	Embedded and Networked Systems
	Ir. Agathe Balayn,	Web Information Systems

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft Delft University of Technology

Preface

This thesis marks the end of my academic career at the Delft University of Technology. As a kid, we would drive past Delft and I would always look at the EEMCS building. I always dreamt of studying there and eventually graduating. Dreams do come true. First, I would like to thank my supervisors, Jie and Agathe, for their excellent guidance and help throughout this journey. Next, I would like to thank my family and friends, who have supported me continually, and were always there for me when I needed them. Continuing with my roommate, the chats, drinks and overall presence was very meaningful to me. Lastly, I would like to thank my mom, Shanita, for the endless love, being someone to look up to, always believing in me, and pushing me to be the best version of myself.

Ashay Somai
Rijswijk, December 2021

Abstract

This thesis looks at how to characterize weaknesses in machine learning models that are used for detecting privacy-sensitive data in images with the help of crowdsourcing. Before we can come up with a method to achieve a goal, we first need to make clear what we consider privacy-sensitive data. We took the General Data Protection Regulation (GDPR) as a starting point, and performed a crowdsourcing task to see how workers interpret this regulation. Interpreting legal texts can be difficult, there is room for interpretation and the perception of a legal text can change over time. Therefore, we need to take the input of the crowd, next to our own input, to operationalize this regulation to use in this context. Next, we took a machine learning model for detecting privacy-sensitive data in images in order to retrieve saliency maps, which helps us with explaining the inner-working of the model. Subsequently, the saliency maps are inspected through a crowdsourcing task, with the established privacy definition, to find out the strengths and weaknesses. From the results, we see that crowd workers can be efficiently used to find the strengths and weaknesses of a machine learning model, while keeping the privacy definition in mind. Workers are able to consistently apply their views about privacy across different images, whilst also increasing the trust people have in the machine learning model. This shows us that we can use crowdsourcing efficiently in a fairly difficult context of privacy, and paves the way for a more sophisticated approach to privacy-sensitive elements in images, and even for contexts other than privacy.

Contents

List of Figures

List of Tables

1

Introduction

The advent of the internet changed countless aspects of our daily lives, including how we learn, communicate, work, think, and define different values, like privacy. One of the areas where privacy is getting a prominent place is object detection in images, for example detecting credit cards or medicine prescriptions. When sharing images, for example on social media, users often overlook these privacy-sensitive elements, even though privacy is an important aspect in our lives, and it should be handled with care. With the introduction of the General Data Protection Regulation (GDPR) in 2018, which provides greater protection and rights to individuals, there are now regulations that describe what privacy-sensitive data is. Next to this, the whole data landscape has been changed, companies and data processors in general need to be more careful with their data and even change their whole chain of operations [4]. Applying the laws written in the GDPR to images is not straightforward, as the GDPR is not specifically written for images. It contains information about what is considered personal data and how it should be handled. How this is applied to images is therefore open for interpretation, and an unambiguous answer does not yet exist. In Orekondy et al. [23] they create a dataset which contain annotations of privacy-sensitive data. They base their definition of privacy-sensitive elements in images off several laws, including the precursor of the GDPR and the US Privacy Act of 1974, but aside from mentioning these laws, a clear explanation on how they decide what is considered privacy-sensitive data lacks.

There are two main approaches for privacy detection in images, the first being machine learning, and the second being with the use of humans, with each having their own benefits. Machine learning approaches are quite fast and can provide users with immediate information about their image, whether it contains privacy-sensitive information or not. An approach that is fast is particularly useful in situations where people are about to post an image on social media, and are directly able to see if there are any privacy-sensitive elements in the picture. After reviewing this judgement, they can proceed with posting the image, or abort it. Next to this, machine learning approaches give an objective and consistent judgement.

With machine learning (and with humans), errors are inevitable. However, as we deal with privacy, these errors could be disastrous, breaking the laws with possible repercussions, and therefore should be reduced to a minimum. For example, a blind person that uses VizWiz [8] for visual questions. With this platform, the users send a picture almost immediately to crowd workers, to get an answer to a question the users pose. However, when this picture contains any privacy-sensitive elements, like a credit card or a medicine prescription, that are missed by an automated privacy-detection tool, the crowd workers can potentially misuse this information. Thinking about reducing errors in machine learning, gives us some options. The first one could be: increase the size of the dataset. If we follow our intuition, this would mean that the machine learning model would see more examples, which would lead to a higher accuracy. However, there are at least two drawbacks: increasing the size of the dataset does not necessarily lead to a higher accuracy and if it did, acquiring a larger dataset will cost more, monetary or time-wise. The second option is to tweak the algorithm in such a way that the accuracy increases. The major drawback is that this might be the (almost) optimal solution for one specific dataset, and does not necessarily hold for other datasets, which, in turn, need other refinements of the algorithm. Robustness is an important aspect that we need.

On the other side, human approaches (crowdsourcing) have a low error rate, which is important when dealing with privacy. Any errors that are made can be disastrous and should be avoided entirely. It is hard for a single person to catch all the privacy-sensitive elements, so with crowdsourcing multiple people go over the same instance, which reduces the error rate. Another benefit human approaches have, is that they are good at looking at the context of an image and adapt their judgement based on it, which some instances regarding privacy might require. This goes for elements that are safe in isolation, but become privacy-sensitive in a particular context, and it goes the other way as well, elements that are safe when incorporating the context.

Using solely humans also has its drawbacks. First and foremost: they are expensive. Especially when dealing with large amounts of data. Next to this, humans, in our case crowd workers, have varying backgrounds and might think differently about various subjects. When deciding whether an image contains personal data or not, these differences might have a large impact and will lead to a large variety of answers. One might be more comfortable sharing their data on the internet opposed to someone else, who tries their best to not leave any trace of them on the internet. For example, people from cultures with high individualism believe in the right of a private life, and are thus more concerned about the privacy of their personal information than people from more collective cultures [10].

From what we can see above, using only machine learning, or only humans for detecting privacy-sensitive elements has its benefits, while they also have their drawbacks. Using humans to fill in the gap where machine learning falls short, and thereby overcoming the weaknesses of machine learning, is a possible solution. Therefore, we argue that a combination of both machine learning and humans might be suitable to tackle this problem, and this leads us to our hypothesis: "Including humans in the

detection of privacy-sensitive data in images will positively impact this process."

1.1. Problem Statement

In this thesis, we will characterize the strengths and weaknesses of a machine learning model for detecting privacy-sensitive data in images and overcoming these weaknesses by incorporating humans in the process. The idea is that we use a machine learning model to identify privacy-sensitive elements in images. We assume that the model is not capable of doing a flawless detection of all the different personal elements in the images, and might be better in detecting some classes, but worse in other classes. After we have identified these particular classes, we can continue with improve the "bad" classes with the help of humans, and further improving the "good classes". This leads us to our research question: **"To what extent can we use humans efficiently to increase the detection of privacy-sensitive elements in pictures?"**

The first challenge that we have to deal with is the interpretation of the GDPR for images. As mentioned earlier, there does not yet exist a clear and concise explanation and motivation of applying the GDPR to images. The thing that we do have, as a starting point, is the dataset provided by Orekondy et al. [23]. Next to this, it is also important to capture the interpretation of the GDPR by humans, as we do not want to create our version of the truth, but one that is widely accepted. So, together with inputs from the crowd and ourselves, we need to operationalize what can be considered privacy-sensitive in images, before we can continue with the rest of the research.

The second challenge is how do we characterize strengths and weaknesses of a machine learning model in this context. The usual metrics, such as accuracy and precision, does not reflect the inner workings of a model, in particular where they look in the image itself when making a decision. Therefore, we need to come up with a method to execute this characterization.

Revisiting the research question, we further divide this into sub-questions:

1. RQ1: What is the state-of-the-art regarding the notion of privacy and privacy detection in images?

2. RQ2: How do we operationalize privacy-sensitive elements in the context of images with the help of humans?

3. RQ3: How do we characterize the strengths and weaknesses of a machine learning model with the help of humans for detecting privacy-sensitive elements in images?

For RQ1 we will look into what is currently researched, what solutions are out there and what limitations they have. For RQ2 we formulate a clear and concise definition of what is considered privacy-sensitive in images. This will include an analysis of the GDPR and how humans apply these laws. For RQ3 we will train a model and identify the weaknesses (and strengths) of said model and come up with a taxonomy of the different classes.

1.2. Thesis Contribution

- C1: A comprehensive analysis of current research about the notion of privacy and privacy detection in images (RQ1)
- C2: An assessment of relevant privacy laws and interpretability of laws. (RQ2)
- C3: A study on how humans annotate privacy-sensitive elements in images given a definition of privacy. (RQ2)
- C4: A method to characterize the strengths and weaknesses of a machine learning model for detecting personal data in images (RQ3)

1.3. Research Outline

In chapter 2, we will look at the background and related work. Next, in chapter 3 we will take a look at the various questions we have and discuss the methodology used in order to get an answer to these questions. Following, in chapter 4 we will take a look at the different experiments that we run, how they are set up and what the results are that we obtain from these experiments. Lastly, in chapter 5, we discuss our findings, what the limitations are, the implications of our work, and provide an outlook of what future work can be done.

2

Background and Related Work

In this chapter, we will go through the relevant previous works and state-of-the-art regarding detecting privacy-sensitive data in images. First, we will look at how people define privacy in this setting and how this is applied to their approaches/solutions. Following, we have two approaches for detecting privacy-sensitive data in images, those are a machine learning approach, where the detection is done automatically, and a crowdsourcing approach, where the detection is done manually, with both having their own benefits and limitations. Lastly, we take a look at how recent works tend to characterize weaknesses of machine learning methods for object detection.

2.1. Notion of Privacy

The right of privacy is part of the 1950 European Convention on Human Rights, which states, "Everyone has the right to respect for his private and family life, his home, and his correspondence."[1], which forms the basis for the protection of individuals through legislation by the European Union. Since a few years, we have the General Data Protection Regulation (GDPR), that dictates how one should handle the processing and storing of data in the European Union. There was an increasing concern on how companies and institutions handled data, with these companies and institutions often asking (and storing) more information about users/people than necessary. This would lead to situations where people's data would be sold or given to third parties, such as the Facebook-Cambridge Analytica scandal [9]. Next to this, people's data would be stored for a long time, without the option to get this data removed or altered when it was not up-to-date. With the introduction of the GDPR, the whole data landscape started looking less like the wild west and companies and institutions had a strict set of rules that they need to adhere to, such as informing the user beforehand on what data they collect, collecting only the data necessary, process data fairly and store data for a limited time. This is complemented with several rights that the users have, such as the right to be forgotten and the right to rectification.

[1] `https://gdpr.eu/what-is-gdpr/`

As the GDPR is there for data in general, it does not specify how this works for specific data types, such as images and videos. A couple of papers, that deal with these types of data and how to deal with privacy-sensitive elements, have tried to answer this question, and some even came up with a definition on what is considered privacy-sensitive data in a specific type of data, such as in Pandit et al. [25]. In Pandit et al. [25], the authors go through different tools for consent management and tools that keep track of processed data, and they see a clear difference between vocabularies used to describe the same data principles. An agreement on these vocabularies provides sufficient transparency and provides interoperability between tools. They go on with creating the different classes and subclasses, such that it captures the meaning and aligns with regulations such as the GDPR. In Orekondy et al. [23], they use machine learning in order to redact images, they remove privacy-sensitive information, while still preserving intelligibility. But before they address this challenge, they look at suitable datasets for their problem and augment/enhance the dataset, such that it better fits their purpose. They start off with the VISPR dataset, as introduced in [24]. The VISPR dataset consists of "22k images that allows the study of privacy relevant attributes in images and the training of automatic recognizers" [24]. There are 68 different privacy attributes, that are compiled by looking at multiple relevant privacy sources, such as *Guide to protecting the confidentiality of personally identifiable information* [22], the EU Data Protection Directive 95/46/EC [1] and the US Privacy Act of 1974 [3]. The statistics of the VISPR dataset can be found in table 2.1. The authors of [23] directly build upon this work by filtering the dataset such that it fits their goal of removing privacy-sensitive information while still preserving privacy. They start off by removing all the images that are labelled as "safe", which reduces the size of the dataset with 10k. The next step is selecting only the pictures that contain at most five people, as they wanted to reduce the annotation cost while still retaining non-person areas in the pictures. As for the privacy attributes, they state how they are compiled, the same as in [24], but aside from mentioning the sources, there are no specific passages mentioned and no rationale or motivation provided. So the actual application of the relevant laws is missing and could be done in a better way. Next to this, both papers rely on the EU Data Protection Directive 95/46/EC, that has been repealed and replaced by the GDPR.

Interpretation is, according to the Legal Dictionary [2], "The art or process of determining the intended meaning of a written document, such as a constitution, statute, contract, deed, or will", and is still the subject of many discussions, when applying existing laws, or when new laws arises, and is considered fundamental to the practice of law. Whenever the meaning of the words in the text is vague and the intent of it can not be deduced, it might be interpreted in different ways. In Barak [7] and Greenberg [16], the authors try to systematically approach this problem, and try to come up with a suitable approach or system to deal with legal interpretation. Greenberg [16] concludes that, despite a large literature, there exists still a large unclarity on what the different approaches amount to.

What we can see is that there are many works regarding this problem, however these works are mostly from a legal interpreter's point of view, and there is still no

Split	All	Train	Val	Test
Images	22,167	10,000	4,167	8000
Labels	115,742	51,799	22,026	41,917
Avg Labels/Image	5.22	5.18	5.29	5.24
Max Images/Label	10,460	4,710	1,969	3,781
Min Image/Label	44	20	7	12

Table 2.1: VISPR Dataset Statistics [24]

consensus (which is perhaps not possible), but there is a lack of views from "ordinary" people. So it is interesting to see how "ordinary" people interpret a certain passage of a law, whether they can reach a consensus among themselves, and if their performance is comparable to that of a legal interpreter.

2.2. Machine Learning approach for privacy detection

As mentioned in the previous section, Orekondy et al. [23] uses machine learning in order to redact images, removing privacy-sensitive information while still preserving intelligibility, a brief idea of their work can be seen in figure 2.1. They create an extension of the VISPR dataset [24], where they remove the privacy attributes that are linked to the entire image, such as sports and religion (and thus the whole image needs to be redacted, so does not preserve intelligibility), attributes that are extremely tedious to annotate, such as political and general opinion, because of their strong co-occurrence with crowd-scenes and attributes that have too few examples to incorporate in the eventual model. Next to this, they also merge a few attributes when they are present in the dataset as complete and partial version (partial face and complete face into face) and when the attributes localize to the same region (race, skin color, gender into person). With this new set of attributes they proceed with annotating the dataset, with the following results: "With an annotation effort of ~800 hours concentrated over four months with five annotators (excluding the authors), we propose the first sizable pixel-labeled privacy dataset of 8,473 images annotated with ~47.6k instances using 24 privacy attributes." Next, they commence with a user-study in order to find out what the effect is of the amount of redaction in an image on the privacy and utility of said image. With the results, they found an optimal value for the trade-off between privacy and utility. Then they start with automating the detection and redaction of the different attributes, and qualitative examples can be found in figure 2.2. With their method, they achieve a performance of 83% when compared to the ground-truth.

The precursor to the paper presented in the previous paragraph is [24]. This paper uses machine learning in order to provide users, who upload images on various social media platforms, with a tool, that gives images a privacy score based on their own preferences. They conduct a user study, in order to get a clearer view on what is perceived as privacy-sensitive by the users. The users are presented with test images and asked to assess the images according to their own privacy preferences, on a scale from 1 to 5. One of the authors' findings is that people have clear privacy preferences, but

Figure 2.1: Basic pipeline of the work presented in Orekondy et al. [23]

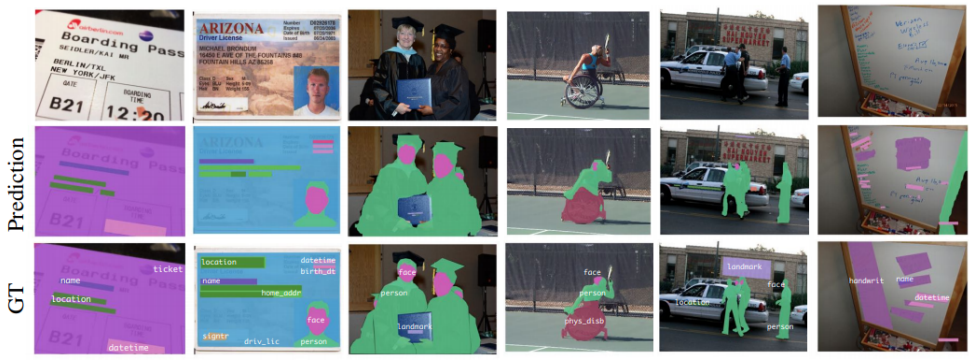


Figure 2.2: Qualitative examples from the method used in [23]

they do not tend to follow these preferences when uploading their own images. Next to the dataset they create, as described in section 2.1, they propose a solution, where they use the different features of the dataset as input, to make a CNN that is connected to a linear SVM. They present two results, the first one being with the user preference included, and the second one without. They show that the machine prediction of privacy risks on images has the edge over human judgement, assuming that humans are not necessarily following their own privacy preference.

2.3. Crowdsourcing approach for privacy detection

Next to machine learning, there are also crowdsourcing approaches for privacy detection. In these approaches, humans play a crucial role, and particularly many humans.

Often, people use crowdsourcing to do repeating work that takes up a lot of time if done by an individual. Next to this, it is cheaper to use crowdsourcing instead of hiring experts for a task, partly due to the sheer size of the tasks [18]. Aside from quantitative tasks, crowd workers can also be used for qualitative tasks. Crowd workers all come from around the worlds, and therefore have different backgrounds, age, gender et cetera [29]. With this variety of people, you can essentially find out how different people think about certain topics, for example, someone from the western world might prefer coffee to tea, whereas people from the eastern world prefer tea to coffee. This is also applicable to privacy, in some areas of the world people share more, whereas people in other parts of the world are more reserved. Aside from this global differences, differences can also occur on a smaller scale.

Starting with Alshaibani et al. [5], the authors begin with stating that despite the recent improvements in automated face detection, humans still perform better than machines. In contexts where privacy is important, this edge is crucial. Therefore, the authors propose an approach based on human perception, however, the authors say that an ultimate solution combines both human perception and automated detection. The approach is that an image is first filtered using a median filter, to make the faces in the images unrecognizable. Crowd workers are asked to place an ellipse on areas where a face might be. The results of the workers are combined, and the resulting area is saved. In the next step, the face areas retrieved from the workers remain unchanged, but the rest of the image is now filtered with a median filter with a lower k-size, possibly revealing more faces that can be detected by the workers. This is done iteratively, until all the faces are redacted. They conduct a study to decide which boundaries of k-size to use (which impacts the blur-level), as well as the number of steps needed, in order to keep the probability of faces detected high (P>=0.98) and the probability of faces recognized low (P<=0.02). The results show that 98.7% of the faces were detected, showing improvement over Microsoft Azure's face detection system that detected 87.5% of the faces. As for the identification, workers were able to correctly identify 0.83% of the faces (seven workers on seven distinct images). One major limitation of this work is that it is focused only on faces, whereas there exists many privacy-sensitive elements, with different granularities. Incorporating these different elements in this approach might cause the need for more steps (to look for different granularities) and increase the time a worker needs to fulfil the task (and thus increasing the overall time and cost to assess an image).

A paper that also has an iterative approach, but looks at more than just faces, is CrowdMask [19]. In this paper, the authors propose a pyramid segmentation workflow, in order to mask privacy-sensitive elements in images. Workers are shown small segments of the image and are asked to look for any (parts of) elements that might be privacy-sensitive. These elements are masked once there is a consensus between workers, and continues to the next step, where a larger segment is shown - with the masks retrieved from the previous step - in order to find elements of different granularities. This is continued until the whole image is assessed. Besides this, because they are aware that the tasks might be time-consuming (as the number of steps can be

set freely) and costly, they propose two things. The first one is a segmentation optimization function, where you have to fill in a parameter budget, in order to know what the most efficient segmentation is (i.e., in how many segments should the image be divided into, how many steps are needed). Next to this, they also show that privacy-sensitive elements in segments can be predicted, for example a face can span multiple segments, and you can say, with a high certainty, that the adjacent segment also contains a part of a face, if the segment you are looking at contains a part of a face, and this reduces the workers needed for a consensus.

Even though the chances are slim, there still exists a possibility where the smallest segment, used in the paper in the previous paragraph, contains personal elements in its entirety. This is visible to the crowd workers, when dealing with an automated system where the images are not assessed beforehand. This might be harmful and is preferably avoided at all costs. In Human OCR [21], they look at how a crowdsourcing task can be fulfilled by workers without revealing privacy-sensitive information or information that is not necessary for the task. They have a specific use-case, where they try to digitalize handwritten forms. Instead of showing the whole form (i.e., showing all the information that is written on it), they only show the cells that contain text, with the corresponding labels. The idea behind this is that workers can not do anything with individual pieces of the form. This requires the form to be pre-processed, to retrieve the cells that contain text. In this pre-processing step, workers are asked to view a blank form and select the cells that need to be filled in, together with the corresponding label (i.e., name, date-of-birth, medical conditions). This paper sheds a different light on detecting (and masking) privacy-sensitive information. Instead of trying to rigidly find all the different elements and masking them one by one, they show workers the entire elements, but without revealing the surrounding information. These elements, on its own, do not necessarily carry privacy-sensitive information, but in unison with the surrounding information, this might be the case. Therefore, it is useful to look at those privacy-sensitive elements, and see whether they can be considered "safe", when it is the only element visible.

One paper that does not necessarily tackle the privacy problem, but still could be relevant is Das et al. [11]. This paper tries to give a solution to a problem that moderators of for example social media platforms have: they are exposed to harmful content which could cause lasting psychological damage. Automated detection is difficult due to the high accuracy requirements, costs of errors, and nuanced rules for acceptable content. So ultimately, moderation systems require some level of human labour in order to make difficult or final judgement calls. Their approach is to blur the image, such that the image is not fully visible to the moderators, which would be crowd workers in the experimental setup. They have different levels of blur: no blur, medium blur, strong blur, slider, click and hover. As this is their approach, they put it to the test and see how the workers respond to the different approaches, such as the positive/negative experience and emotional exhaustion. Their results show that different approaches have different impacts on the workers. In our case, as we want to insert humans in the privacy detection process, we need to have an idea of what the workers

will see and prevent them for seeing potentially harmful content.

In Han et al. [17], the authors try to capture the privacy attitudes of social media users and see whether they change in certain scenarios. They compiled a list of attributes that might reveal private information on Instagram profiles, such as birthday, hometown and real name. Their experiment began with asking for some information about the workers, such as name and gender, and also the average length and frequency of their Instagram uses. Subsequently, the attributes are presented to the worker, and they are asked to rate them how concerned they would be if this attribute would be revealed on their profile. In the next step, the workers are shown real Instagram profiles, which contain at least one of the attributes and all the profiles shown combined cover all the attributes, and are asked to look for any revealing information and how concerned they would be if that type of information is shown on their own profile. After a 5-minute break, they are again asked to assess the individual attributes and rate them how concerned they would be if this attribute would be revealed on their profile. From the results they see that for some attributes there is a large shift in how the workers rate them, some attributes are only deemed concerning after the workers see an example of that attribute on a real profile, whereas other attributes decrease in concern. This work also confirms the existence of the privacy paradox [28], "A discrepancy between subjective perceptions and objective selections could imply the case where a user is concerned about a certain privacy item being exposed, but does not illustrate a corresponding action (such as removing or masking the item)"[17].

2.4. Characterizing Predictions in Machine Learning using Humans

Machine learning approaches are often evaluated by looking at several metrics, such as precision, recall or accuracy. However, in the context of privacy, a (relatively) good accuracy score, but less than 100%, is still not good enough, as the misses might have disastrous impacts. Therefore, people might be hesitant towards using the model in real-world applications [29] and if they see the model making any mistakes, they avoid using it at all [13]. Therefore, to increase the trust of these models, there is a need to get to know the inner workings of a model or explain the results of a model, in order to find the shortcomings and act on them. These inner workings or results are hard for people to understand, and there is now an increase in research in how to make these model human-interpretable [29].

In Ribeiro et al. [26], they propose an algorithm called LIME, that explains the prediction of any classifier. They argue that the role of humans is overlooked in recent advances in machine learning, but in the end, humans are the main consumers, directly or indirectly, of these classifiers. When a human does not trust the model or prediction, they will not use it at all, is what they say, and is their primary motivation for this technique that they propose. They aim to gain this trust from the users, by showing them individual predictions and explanations, next to the evaluation metrics. One example that they show is about object detection, where they explain the prediction of Google's pre-trained Inception neural network [27] on an image of a dog playing a guitar, which

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 2.3: Example of explaining individual predictions [26]

can be seen in figure 2.3. In this image, three classes are predicted, *Electric guitar*, *Acoustic guitar* and *Labrador*, and the areas that are responsible for the prediction are only shown. Even though one class is predicted wrongly, this enhances the trust of the user in the model, as for the correct classes the model looks at the right areas, and for the wrong class, the model is not too far off. This also helps with assessing predictions that are completely off, they might require tweaking the model or more samples in the dataset.

Zintgraf et al. [30] takes visualizing which areas of an image contribute to a prediction to the next level, where they also show areas that contribute against the prediction. This offers more explainability, helps with research on the networks, and the usability and acceptance of the model. This is especially important in the context of privacy, as it is useful to know why a certain element is not considered privacy-sensitive. They use a multivariate analysis, where they estimate the relevance of a feature of the model by measuring how the prediction changes when this feature is removed. In figure 2.4, the prediction made is "cockatoo", which can be seen on the original image on the left. The corresponding visualization shows red areas, which are in favour, and blue areas, which are against. The facial features point the model in the direction of a cockatoo, however the other parts are pointing towards a white wolf, which is the second-best scoring class.

Making the inner-workings and results of a model interpretable for humans, allows them to act deduce the strengths and weaknesses of the model. In Ribeiro et al. [26], after obtaining the individual predictions and explanations, workers are asked to inspect these and assess whether the explanation makes sense or not. In the case of wrong explanations, you can assess multiple instances of this particular prediction, and conclude whether the model is performing well or not, creating some sort of taxonomy of classes and the overall judgement. With the approach proposed by Zintgraf et al. [30], you can go further and see which classes the model chooses between, and recognize patterns of classes that the model has trouble separating, further increasing the knowledge you gain about the model.

What we see from these works is that characterizing the predictions of a machine learning model has several benefits, such as finding the shortcomings of a model and increasing the trust humans have in these models. One of the limitations that we face

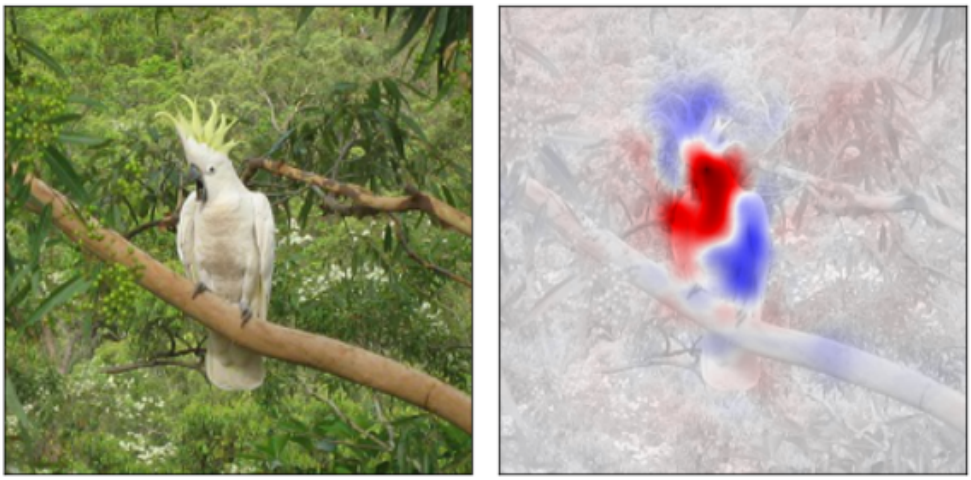


Figure 2.4: The original image and resulting visualization of a prediction [30]

in machine learning approaches for the detection of privacy-sensitive elements is that the accuracy is not good enough to solely use these approaches. Therefore, finding the shortcomings, by looking at the inner-workings by humans, helps with increasing the performance, whilst also increasing the trust of humans in the model. The increased trust helps with the hesitancy humans have [29] and prevents not using the model at all [13]. Thus, it is interesting to see how these approaches would fare in the context of privacy, where the humans have to assess the inner-workings and predictions according to a given privacy definition.

3

Approach

In this chapter, we will start off with a study on the GDPR, what it entails, how this can be applied to images and how it can be interpreted in different ways. Aside from our own perspective on this, we want to know whether this is generally accepted by others, and we do not want to create our own version of the truth. This will be done through a crowdsourcing task, where the aim is for the crowd workers to apply the GDPR to images, without any influence from our side. Combining these, the operationalization of privacy-sensitive elements for images that is generally accepted, and this approach will be evaluated in the next chapter.

Next off, we look at existing visual privacy datasets that are used for the detection of privacy-sensitive elements, and use this in combination with a pre-trained machine learning model, which will give us a simple privacy detector for images. Using the operationalization of the privacy-sensitive elements obtained earlier and with the use of saliency maps, we should get a clear overview of what the model is capable of in terms of strengths and weaknesses. See figure 3.1 for a diagram of the workflow.

3.1. Privacy Definition

Before we can start with looking at ML models and how they identify privacy-sensitive elements in images, we first need to clarify what we define as "privacy-sensitive". We consider this operationalization crucial, as there are many discrepancies on what is considered privacy-sensitive by different people.

One thing that we can use and also provides us a solid basis for our operationalization of privacy-sensitive elements is the General Data Protection Regulation (GDPR). The GDPR is a relatively young regulation, introduced by the European Union, to ""harmonize" data privacy laws across all of its members countries as well as providing greater protection and rights to individuals"[1]. However, the GDPR does not have anything specific regarding images, it is mostly about data in general. Therefore, extend-

[1] https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018

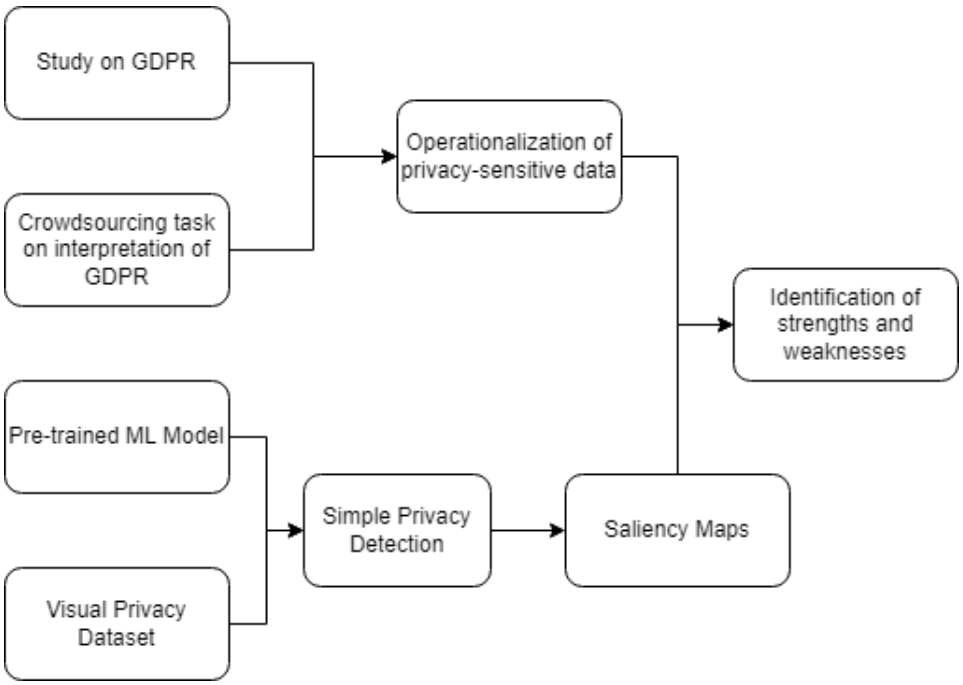


Figure 3.1: A diagram of the workflow

ing the GDPR for images leaves a lot of room for interpretation. Along with the general vagueness of laws, this gives us various interpretations of the same laws.

Looking at how the previous work has dealt with this problem, we see that Orekondy et al. [23] has created a dataset of images, with privacy-sensitive annotations. The dataset contains a set of labels, that the authors consider privacy-sensitive. These labels are selected from an existing list of labels introduced in Orekondy et al. [24], with several privacy laws in mind, such as the GDPR and the US Privacy Act of 1974. So they did not create a list of labels from the privacy laws itself, but rather applied the laws to the pre-existing list and selected the applicable labels. However, the criteria and rationale for selecting labels that are privacy-sensitive lacks.

If we take a look at the list of privacy-sensitive labels used by Orekondy et al. [23] and consider the definition of privacy-sensitive data given by the GDPR as well, then there are some labels that, on their own or in certain combinations, to the best of our understanding, do not pose a threat to the privacy of an individual. Considering the definition given by the GDPR: "privacy-sensitive data are any information which are related to an identified or identifiable natural person. The data subjects are identifiable if they can be directly or indirectly identified, especially by reference to an identifier such as a name, an identification number, location data, an online identifier or one of several special characteristics, which expresses the physical, physiological, genetic, mental, commercial, cultural or social identity of these natural persons." and the data in table 3.1. From this table, it is clear that the identifying information is the *Employee no.*. The city they live in and whether they work from home or not is linked to this

Employee no.	City	Works from home
1234	The Hague	No
2468	Rotterdam	No
3690	Groningen	Yes
4848	Enschede	Yes

Table 3.1: Example of company data on their employees, where they live and if they work from home.

Employee no.	City	Works from home
	The Hague	No
	Rotterdam	No
	Groningen	Yes
	Enschede	Yes

Table 3.2: Example of company data on employees, but redacted such that the *Employee no.* is not visible.

identifier and is therefore also considered privacy-sensitive data. Once we redact the column *Employee no.*, as we can see in table 3.2, then there exists no identifier that identifies a natural person. This means that the columns *City* and *Works from home* are not considered privacy-sensitive data and are even considered safe, opposed to the example with *Employee no.*.

One possible explanation for the label choices of the authors could be that they have a 'greedy' approach. The label itself might not be considered privacy-sensitive, but in certain combinations with other labels, it might become privacy-sensitive, and therefore included in the list. In the case of privacy, then you do not want to take risks with labels, unless you are certain that it is safe to include it. In this context, the 'greedy' approach makes sense. So there is a need for a clear operationalization of what is privacy-sensitive data in images. Going through all the labels one by one and checking every combination myself might be a suitable option to obtain some set of rules on in what situations a label can be considered privacy-sensitive data, but this might be not widely accepted.

Therefore, we propose to conduct a crowdsourcing task, with the goal to capture the subjectivity of the judgement of what is considered privacy-sensitive data or not. The general idea is to give the crowd workers as little information as possible, in order to not influence them in any way. They are given the definition of privacy-sensitive data by the GDPR, and should interpret this and apply it to various images.

3.1.1. Privacy notion task

The goal of this task is to capture the subjectivity of the judgement of what is considered privacy-sensitive data by the crowd workers, following the GDPR. The structure of the task is the following: first, workers have to inspect an image and are asked whether the image contains privacy-sensitive information, according to the definition given by the GDPR. If yes, they are asked to select the elements in the image and assign a label

to it. We further elaborate this in the subsequent paragraphs. The task consists of two parts: the "training" part and the actual task. Before we can make a training part, we first need to discuss the actual task, as the training part depends on the actual task.

In the task, we are going to use an image from the dataset used in Orekondy et al. [23], so that it has a 'ground-truth'. The crowd worker is first asked to thoroughly read the relevant GDPR passage, which is the following: "Personal data are any information which are related to an identified or identifiable natural person. The data subjects are identifiable if they can be directly or indirectly identified, especially by reference to an identifier such as a name, an identification number, location data, an online identifier or one of several special characteristics, which expresses the physical, physiological, genetic, mental, commercial, cultural or social identity of these natural persons. In practice, these also include all data which are or can be assigned to a person in any kind of way. For example, the telephone, credit card or personnel number of a person, account data, number plate, appearance, customer number or address are all personal data.".

After the crowd worker has read this, we assume that the worker is capable of applying this law to an image. The worker is then asked to inspect the image from the dataset, and is subsequently asked if the image contains any privacy-sensitive elements, which is shown in figure 3.2. Based on the answer, the worker has given, we then proceed with the next set of questions. If the worker has answered that the image contains privacy-sensitive elements, they are asked to draw bounding boxes around said elements, and attach a label to the bounding boxes. The overall question can be seen in figure 3.3 and drawing the bounding box and attaching a label in figure 3.4. After all the different privacy-sensitive elements are marked and labelled, the worker is then asked which elements can be used to directly or indirectly identify a natural person, see figure 3.5, and whether the image is still privacy-sensitive when these identifiers are left out, see figure 3.6. The idea behind these two final questions is, to our understanding, that without direct or indirect identifiers, no natural person can be identified, and therefore the image becomes "safe" (see section 3.1). So the expected results are a set of elements that are considered identifiers, individually or in combination with other elements.

In the training task, we start off with a general introduction of the task. Here is discussed what the task entails and what the worker is supposed to do, so starting with carefully inspecting the given image, judging whether the image contains privacy-sensitive elements according to the definition give by the GDPR, and if applicable drawing the bounding boxes and attaching the labels, followed by the identifier questions. The GDPR definition is located in the training task, and the workers have to read it thoroughly here. Next are examples of privacy-sensitive elements, as you can see in the figure 3.7. Lastly, there is a small GIF, that shows the crowd worker how to draw the bounding boxes. The task description and the definition can be viewed within the actual task at any time, as you can see in the top-right corner of figure 3.2.

After we obtain the results, the first thing that we will do is checking if there are

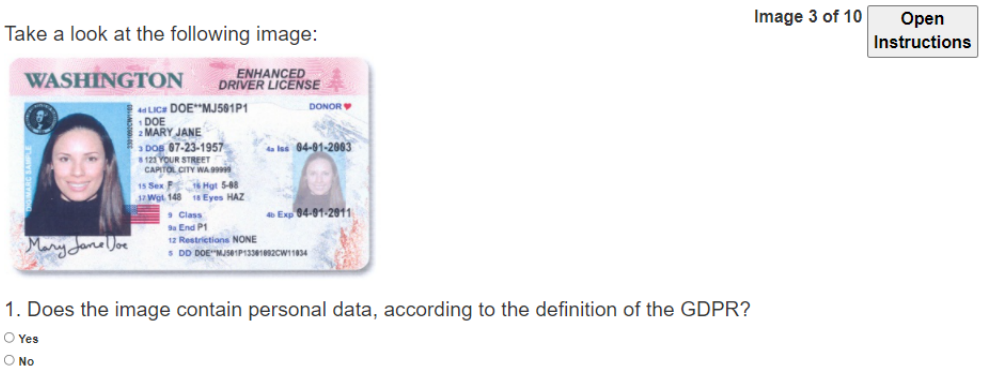


Figure 3.2: The introduction and first question of the privacy notion task.

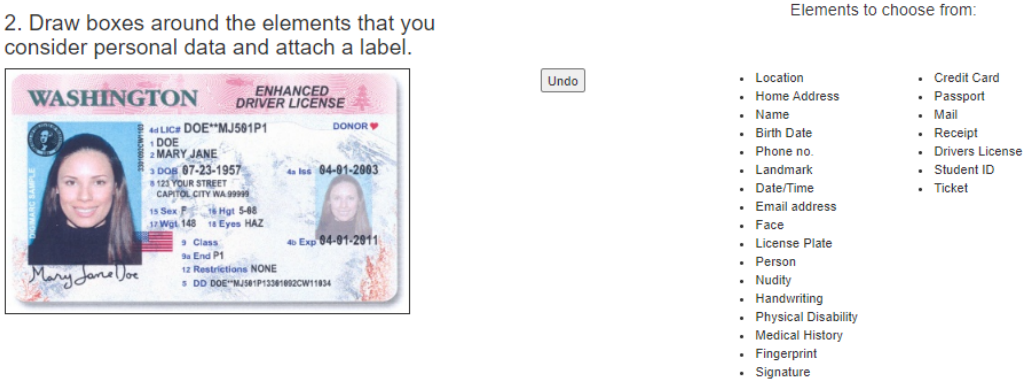


Figure 3.3: The second question of the privacy notion task.

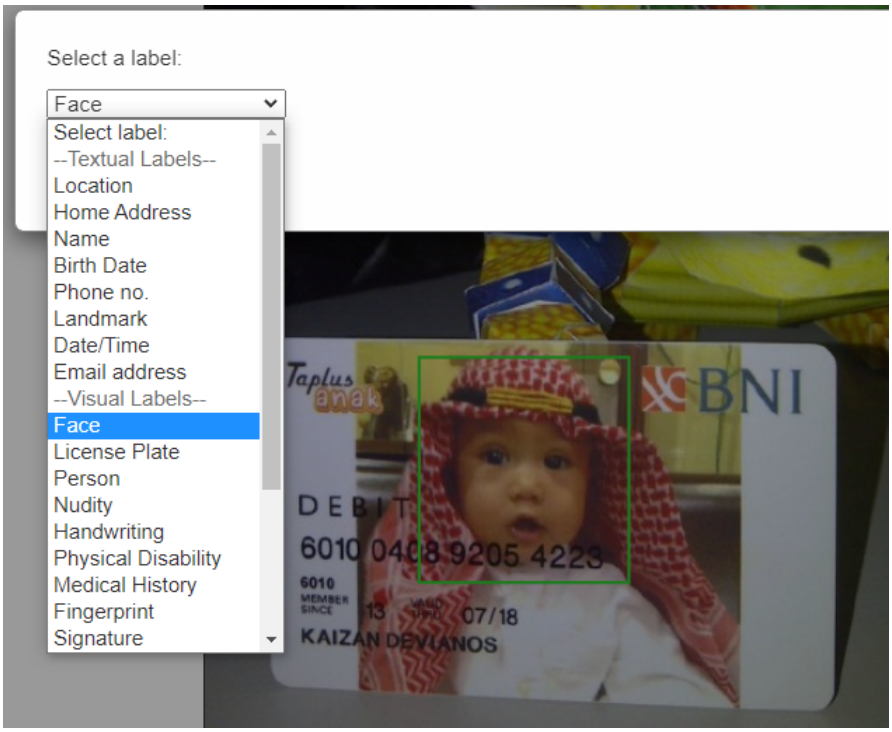


Figure 3.4: Example of the bounding box and attaching a label.

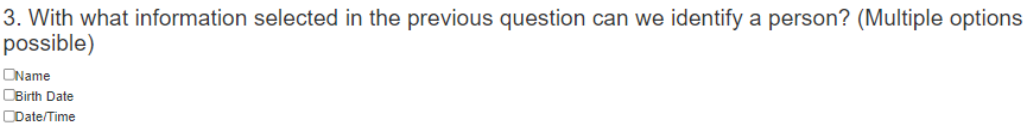


Figure 3.5: The third question of the privacy notion task.

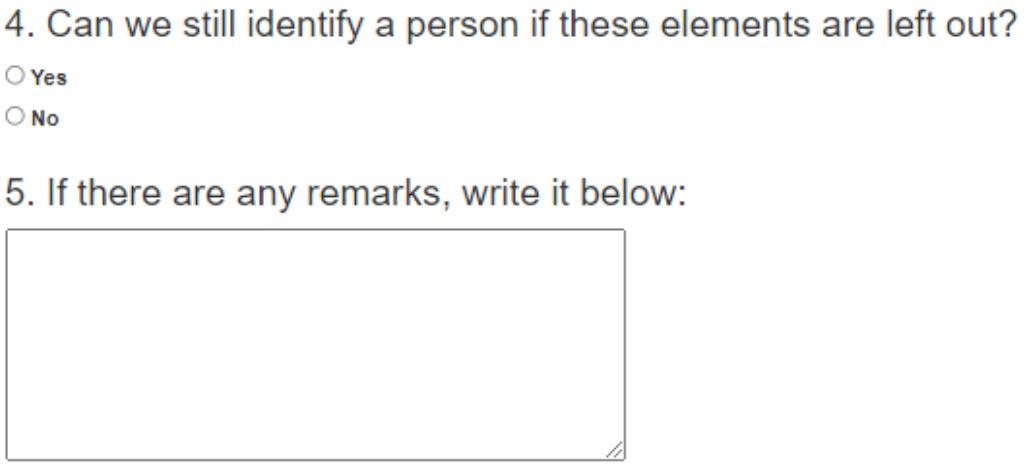


Figure 3.6: The fourth and fifth question of the privacy notion task.

Figure 3.7: Examples of the attributes

Figure 3.8: The original image and its corresponding saliency map for the attribute American Lobster

any workers that did not complete the task as intended, in order to exclude their submissions. Next, we will look at how the responses of an individual worker compare to other workers, whether they are able to reach a consensus or not, with the use of an inter-annotator agreement. Following this, we will look at how the workers' responses compare to the ground-truth, i.e., the annotations in the datasets. This will be analysed, in order to find any patterns that emerge, and an explanation for the pattern will be sought.

3.2. Identifying Weaknesses in ML Models

From the background and related work, we can see that the accuracy when detecting privacy-sensitive elements in images is not that high. This is, however, crucial, as we deal with privacy, and any 'miss' by an ML model might have serious impact. As mentioned in the introduction, there are several ways that you generally would traverse to improve the accuracy of the model, but they do not necessarily work. An alternative would be to use only humans to assess images. But this also has its limitations: too expensive and too slow.

A more preferable solution would be combining both Machine Learning and humans, in order to get, simply said, the best of both worlds. In practice, we opt for running first an ML model and try to identify where the weaknesses lie in the model. We will do this by, for each label or class, look where the model is looking in the image itself. In Balayn et al. [6], they used saliency maps to see which areas of an image are activated for a certain output, as you can see in figure 3.8. By using this, we can correctly identify, to a certain extent, which labels are correctly identified i.e., the correct area is highlighted, which labels are incorrectly identified i.e., the wrong areas are highlighted and everything in between. Eventually, the goal is to retrieve a taxonomy of labels and whether they are correctly identified by the model. This should not be a yes/no, but rather an accuracy score. For the incorrectly identified labels, we then know that those should be augmented by a human.

3.2.1. Implementing the Machine Learning Model

As discussed above, in Balayn et al. [6], they have a machine learning model, for object detection, and they extract saliency maps, in order to see whether the correct areas in the image are activated. Essentially, we want to achieve the same, however there are a few alterations that we need to make, to achieve our goal.

They use TensorFlow and Keras and uses a pre-trained model. This is the InceptionV3, with imagenet weights, with an extra layer at the end for a softmax activation. They take in images of size 75x75, with one annotation per image, and the output therefore is also one annotation per image. Our approach differs, as an image can contain multiple instances of personal data, so an output (and input) of one annotation per image is not sufficient. Therefore, we altered the model, such that the input and output are multiple annotations per image, and changed the activation function of the last layer from softmax to sigmoid. Next to this, we also changed the image size from 75x75 to 256x256, this gives us a better granularity.

As for the evaluation of the model, the authors use in built-in functions to get the accuracy, visualize the predictions and retrieve the confusion matrix. This does not work for us, as we have multiple predictions per image. Therefore, we opt for only using one evaluation method, namely the Jaccard similarity. This is helpful, as it also counts partially correct answers, i.e. two out of four annotations are correctly predicted. Since our goal is to enhance ML models with crowdsourcing and not particularly achieving the highest accuracy possible with an ML model, this evaluation suffices, and we are more interested in retrieving the saliency maps.

Moving on to the saliency maps, the authors retrieve one saliency map per image, as in their use case they only have one annotation. Since we changed this to multiple annotations per images, we also need to alter this method. For each tensor in the model, which corresponds to an attribute that is present in the dataset, we check how high the confidence is. If this confidence is higher than a certain threshold, we include that tensor, and create a saliency map. This threshold can be altered, a lower threshold might give more insights on what the model is looking at in the image, and a higher threshold might give more accurate results.

In the end, we thus retrieve, for each image in the test set, the predictions, and for each individual prediction, the corresponding heatmap, as you can see in figure 3.9.

3.2.2. Correctness of Model task

This task revolves around checking whether the machine learning model looks at the right area in an image when predicting, through saliency maps. There are several design alternatives that we look into, before deciding on one. The first one is a relatively simple task, where the workers are shown a saliency map of an individual attribute and the name of that attribute, and they have to decide whether this saliency map corresponds to the attribute. This is not a complicated task and does not require a large effort from the workers, as it is a yes/no question that they have to answer.

The second alternative is similar to the first one, so the workers are given a saliency map of an individual attribute, but they do not get the name of that attribute. Instead,

Figure 3.9: Example image and the saliency maps for the attribute `Passport`

they have to select the corresponding attribute from a list of all the attributes. This can help us with, for example, attributes that are consistently portrayed wrong in the saliency map, and requires more effort from the workers. On the other hand, due to the large number of possible attributes, this task might be difficult to complete and if we have saliency maps that are outright wrong, i.e. they do not cover any attributes, it can be confusing for the workers.

The third alternative is that the workers are asked to inspect a saliency map, which covers all the attributes detected in that image. The task is then to select all the attributes that are present in the saliency, given the list of all the attributes. This helps us with identifying the correctly portrayed attributes, and also which attributes are missed by the saliency maps. The downside is that the difficulty rises, as we have a lot of attributes, and as we have attributes with different granularities, it might be difficult to spot certain attributes or distinguish them.

The last alternative looks at how well a saliency map covers the attribute. Workers are shown a saliency map of an individual attribute and the name of the attribute, and are asked whether the saliency map covers the entirety of that attribute. This helps us with determining the granularity of instances, such as a saliency map of the attribute `body`, where only the hands and feet are highlighted. This is similar to the first alternative, with an extension.

Looking at all the alternatives, the first and fourth one are the best options, as they are relatively simple and straightforward to do, compared to the other two alternatives that could be confusing for workers. Keeping tasks simple improves the quality of the results, opposed to larger, more difficult tasks [14]. Comparing the first and fourth alternative, we decided that the fourth, with the different granularities, is the best option. This task is relatively easy to do, and gives us more information about the saliency maps, compared to the first alternative. So, we will further elaborate this task and work out all the details contained in this task.

The task therefore shows an image, with a prediction the model made, and then is asked whether the prediction matches the saliency map. It is unlikely that this is a yes/no question, as there are also instances possible where the saliency map only partly covers the prediction or more than the prediction, i.e. different granularities are possible. So, the images will be rated on a scale from one to five, with one being only a small part of the prediction is covered, three being that the saliency maps covers the prediction perfectly, and five being that the saliency map covers the label and more. Next to these five options, there is another option, being that there is no correlation at all. For each image, then a score will be calculated, showing us which classes are consistently predicted right and wrong and how well the predictions cover the attribute, resulting in the identification of the strengths and weaknesses of the model. As we also look at the granularity, additional findings may occur, such as classes predicted as other classes, or a prediction "face" only looks at eyes.

4

Experiments and Results

In this chapter, we will go through the different experiments conducted, by explaining how they are set up, presenting the results accompanied by a thorough analysis and a conclusion that we can make from this analysis. First, we start with the experimental setup of all the experiments, followed by the experiment where we capture the subjectivity of people regarding the notion of privacy. Next, we have the experiment involving the machine learning model, where the goal is to assess is and identify the strengths and weaknesses of this model.

4.1. Experimental Setup

In this section, we will go through the experimental setup of the crowdsourcing tasks, where the specific choices are elaborated and discussed.

4.1.1. Dataset

As mentioned in the previous chapter, the dataset that we will use is the one introduced in Orekondy et al. [23]. This dataset is available for academic and non-commercial use under a Creative Commons Attribution-NonCommercial 4.0 International License. This dataset provides us with a lot of images, that are all annotated. Next to this, they also provide us with explanations of certain design decisions, that help us with understanding how they ended up with certain privacy attributes.

4.1.2. Notion of Privacy Task

For the crowdsourcing task to capture the subjectivity of people regarding the notion of privacy, we ran a small pilot, to see whether the task is done as it is intended by us, whether the task description and questions are clear, etc., get an indication about the task length and if everything of the task works. This pilot is done through Prolific, which is a crowdsourcing platform[1], and the workers are asked to inspect and annotate

[1] https://www.prolific.co/

five images, which are randomly selected. The task length is assumed to be approximately 10 minutes. The amount of workers asked through Prolific is three, and they are paid £5.70, considering the task length. Aside from these workers through Prolific, we also asked two colleagues to perform the task. The crowdsourcing task uses a NodeJS server, which is hosted on Google Cloud, with a Cloud Firestore database.

After the pilot, we continued with the main crowdsourcing task. The approach and design of the task can be found in section 3.1.1. The task consists of 10 images, that cover all attributes at least once. These images are manually picked from the dataset and the distribution of attributes can be found in figure 4.2. This task is annotated by three workers. In total, we have six sets of 10 images, resulting in six different tasks. Workers are not allowed to do multiple tasks, to ensure that they have no previous knowledge of the task and allowing responses from different people. Next to this screening, workers are only allowed to enter the task if English is their first language. As the GDPR, and laws in general, are complex texts, it is wiser to ask workers that are proficient in English, to remove this barrier, as opposed to people whose English might not be that good and will have difficulties understanding these laws.

As for the task length, we initially set it on 10 minutes, with five images in total to be annotated. This, however, turned out to be too long and workers took approximately six minutes to finish the task. With the increase of images to 10, we decided to keep the task length at 10 minutes. Together with three workers per task, we calculated the cost of fulfilling the task. We decided to pay the workers £7.50 per hour, which leads to £3.75 per task with three workers. With the taxes (20%) and the service fee (33%), we end up with a cost of £5.25 per task, and thus with six tasks, a total cost of £31.50.

4.1.3. Correctness of ML Model Task

For the training of the machine learning model, we used the training part of the dataset [23], which contained 3873 images. The dataset is unbalanced in terms of attributes present, as shown in figure 4.1, where you can see that six attributes have a lot of samples, whereas the other ones have a relatively small number of samples. Next to this, we need to pre-process the images, such that it is compatible with the model. The only thing needed, for the images, was a resizing. The images came in various different sizes and were resized to a format of 256 by 256. The other pre-processing step was for the annotations. The annotations of the dataset contained a lot of information that was not necessary for our model. The annotations came in a JSON format, so with a simple Python script, it was possible to extract the necessary information. We started training with a batch size of 64 and 40 epochs. The resulting weights are saved and stored for further training, when necessary. After training, the model is tested against a set of images, resulting in an accuracy of 62.27%. The next step is to retrieve the saliency maps, this is done by loading the weights and putting the test set into the model. From the test set, 100 images are randomly selected and per attribute the saliency map is saved.

As for the hardware, we used a Google Colab instance, with the following specifications:

- GPU: Tesla K80, 12 GB GDDR5 VRAM

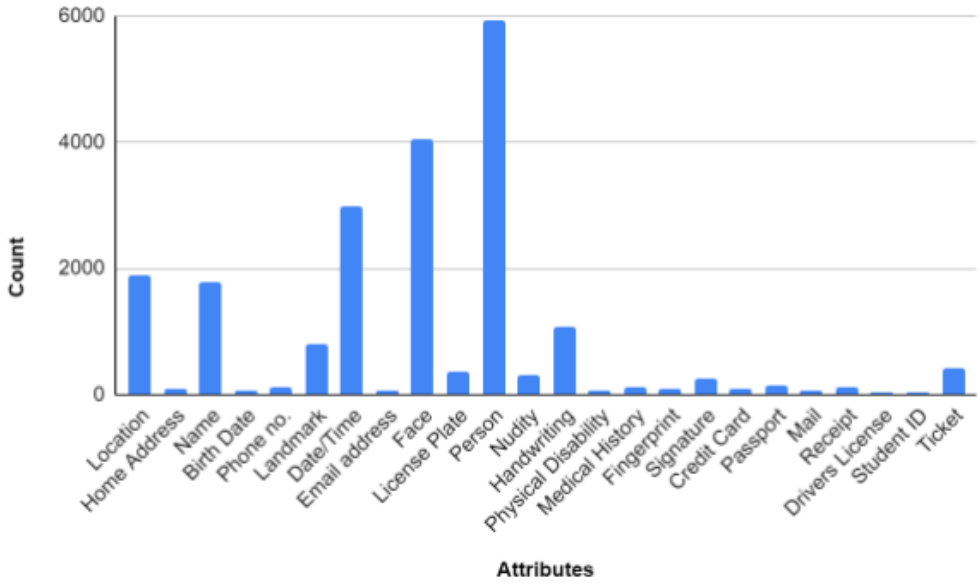


Figure 4.1: Attributes present in training dataset.

- CPU: Intel(R) Xeon(R) CPU @ 2.30GHz
- RAM: 12.6 GB
- Disk: 33 GB

The crowdsourcing task consists of two parts: inspecting the saliency maps and giving them a rating on how well the saliency map covers the given attribute. This rating ranges from 1 to 5, where 1 means that only a small part of the attribute is covered by the saliency map, 5 means that more than the attribute is covered, and 3 would be a perfect fit. Next to this, there is a possibility to give a reasoning for the score, which might give better insights on how someone ended up at giving the particular rating. As for the task length, the number of workers and the monetary compensation, this task will be performed by ourselves, therefore these parameters are redundant.

4.2. The Notion of Privacy

In this section, we discuss the results of the crowdsourcing task as presented in section 3.1.1. Before we ran the task, we conducted a pilot, in order to check the validity of the task itself. In the following subsections, we therefore discuss the pilot first, and subsequently the crowdsourcing task. This section will help us with answering RQ2, which is: How do we conceptualize personal data in the context of images?

4.2.1. Pilot

From the results, we saw that some questions were not formulated properly, leading to wrong answers. These questions were then reformulated in order to remove any possible confusions. Another thing that was interesting is that one crowd worker managed to do as little as possible, not detecting obvious privacy-sensitive elements in the images (opposed to the other participants). Therefore, we added an attention check, to filter out these workers in the post-processing (and not pay them), and added restrictions such as not being able to proceed to the next image without answering key questions. Another bug that we spotted through the pilot was that in the resulting database, one value would always be a `true`. After inspecting the code of the crowdsourcing task, this bug was found and corrected.

The feedback also led to changes in the task itself. These were mostly quality-of-life changes, and not any substantive changes. The first one was regarding the definition of the GDPR on the training page, the suggestion was to make it visually clear what the definition was, either by highlighting it or by putting it between quotation marks. The second one was regarding drawing the bounding boxes, to add an undo button for when they make mistakes. The last major one was to reveal the rest of the questions once the first question is answered 'yes' i.e., the image contains privacy-sensitive elements, and we need to answer more questions. Finally, we also got an indication of the task length, it took around six minutes for participants to finish the task.

Another change that we made, that is independent of the feedback we received, is the number of images used in the task. In the pilot, we used five images, with not every attribute present in the entire task. A better way is to include all the attributes at least once in the set of images, similar to the approach in [17], this reduces the imbalance of attributes present in the images per task. After inspecting the dataset, we concluded that the minimum number of images needed, where all the attributes are present at least once, is 10. The exceptions are attributes that are present in images with many other attributes (mail, ticket, receipt) and attributes that are NSFW (nudity).

4.2.2. Interpretation of the GDPR

After the completion of the task, we first inspected the responses of the workers, in order to find any anomalies. This manual inspection led to the rejection of one worker, who deemed every image as not privacy-sensitive and took way longer than expected to finish the task. There was no objection by the worker, so we can assume that the worker indeed did not correctly perform the task. After processing the results, we took the union of the workers' responses per image and compared them to figure 4.2, which can be seen in figure 4.3. Looking at this figure shows us that, aside from a few obvious disagreements, that the combined answers of the workers follow the ground-truth. To get a closer look at how the workers' responses compare to the ground-truth, we calculated the percentual change, which can be found in table 4.1. From the table you can see that the largest disagreements(+/- 40%) are in the attributes `Birth Date`, `Phone no.`, `Date/Time`, `Person`, `Handwriting` and `Person` (not taking `Receipt` into consideration as explained before).

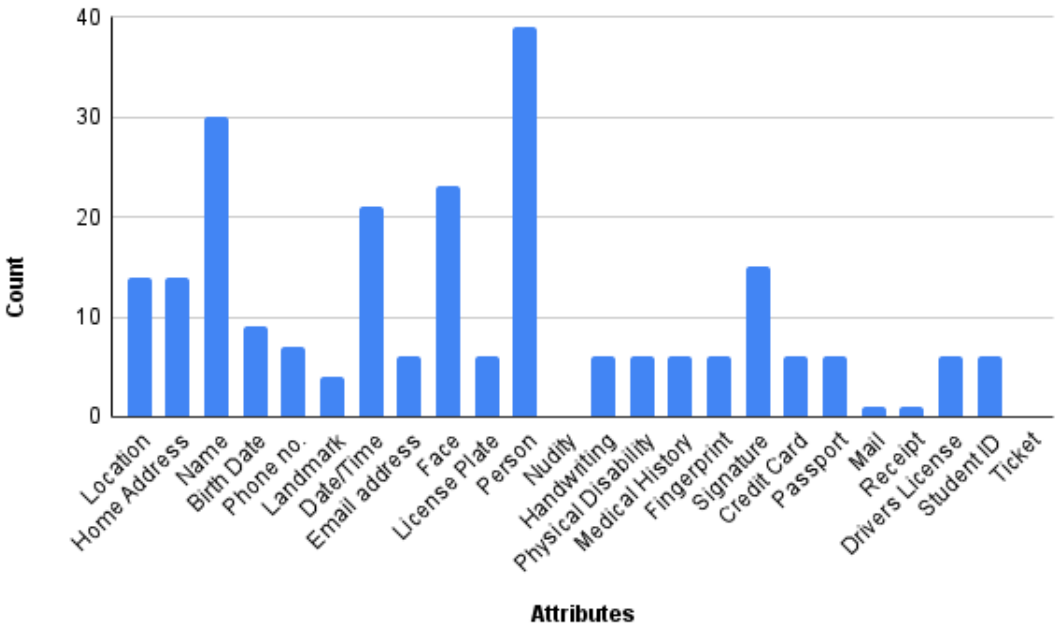


Figure 4.2: The count per attribute present in the task.

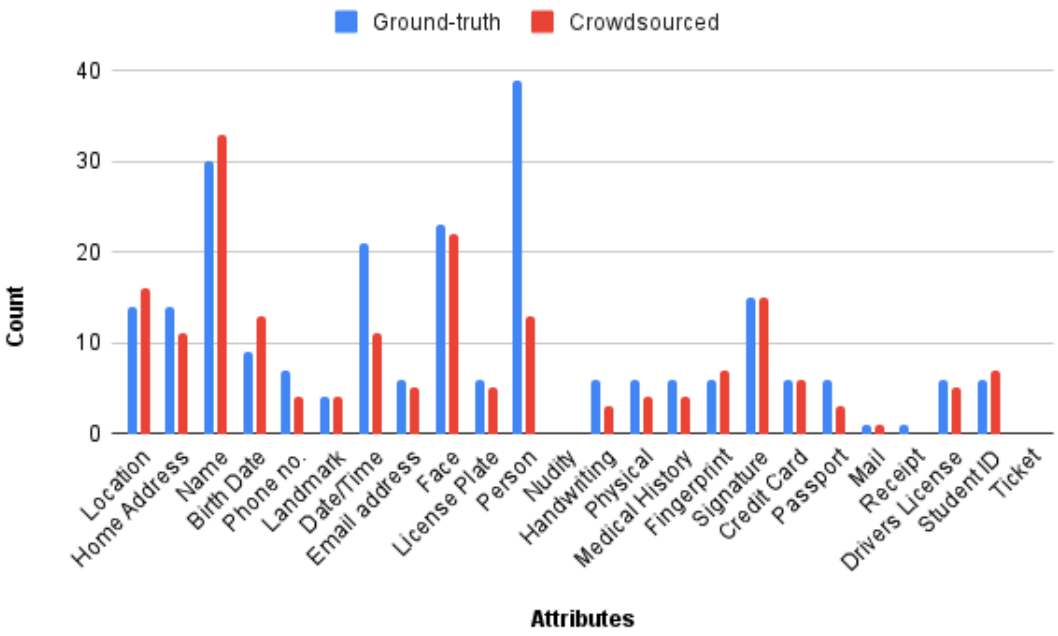


Figure 4.3: The count per attribute by workers compared to the ground-truth.

Attribute	Percentual Change	Attribute	Percentual Change
Location	14,29%	Handwriting	-50,00%
Home Address	-21,43%	Physical Disability	-33,33%
Name	10,00%	Medical History	-33,33%
Birth Date	44,44%	Fingerprint	16,67%
Phone no.	-42,86%	Signature	0,00%
Landmark	0,00%	Credit Card	0,00%
Date/Time	-47,62%	Passport	-50,00%
Email address	-16,67%	Mail	0,00%
Face	-4,35%	Receipt	-100,00%
License Plate	-16,67%	Driver's License	-16,67%
Person	-66,67%	Student ID	16,67%
Nudity	0,00%	Ticket	0,00%

Table 4.1: Percentual change of the count of attributes.

However, these results come from the union of the workers' responses, so it is also important to look at the agreement between the workers themselves. The common approach is to calculate the inter-annotator agreement, which is a measure of how well annotators can make the same annotation decision on a certain category. As we have multiple annotators (three) per images and a multi-label classification, Krippendorff's Alpha is used to calculate the inter-annotator agreement, using the method described in De Swert [12]. The reliability of annotators is calculated per image, where their annotations are compared to each other, and this is done for each image. Eventually the average is taken, resulting in a reliability score of 0,599. This reliability score indicates a moderate agreement, according to Landis and Koch [20], and is close to substantial agreement. One of the reasons for this relatively low agreement could be that there are too many attributes to choose from in the annotation task, next to that it is a multi-classification task. So, in theory, there are many options to choose from when annotating an image and reaching an agreement might be difficult.

Therefore, we decided to look at the agreements for each attribute individually, in order to find out whether the annotators agree on the presence of a certain attribute, without incorporating the different options into the metric. This inter-annotator agreement is calculated using Fleiss' Kappa [15]. For each attribute, the ratings of the workers are inspected beforehand, and only the images with at least one rating, i.e., a worker thinks the attribute is present in the images, are selected. This, however, leads to only a few samples for the attributes, which impacts the validity of the score. Therefore, we only calculate the Kappa of attributes that have at least five images where this attribute is present, according to the workers. The scores can be found in table 4.2, and for each score the strength of agreement is shown, according to Landis and Koch [20]. As you can see, for most of the attributes, the strength of agreement is either moderate or fair. For three attributes, namely `Location`, `Date/Time` and `Person`, the Kappa is smaller than zero, indicating that the expected agreement between the workers was less than

Attribute	Fleiss' Kappa	Strength of Agreement
Location	-0,33	Poor
Home Address	0,39	Fair
Name	0,47	Moderate
Birth Date	0,38	Fair
Date/Time	-0,33	Poor
Email address	0,2	Slight
Face	0,47	Moderate
License Plate	0,47	Moderate
Person	-0,33	Poor
Fingerprint	0,43	Moderate
Signature	0,29	Fair
Credit Card	0,33	Fair
Driver's License	0,73	Substantial
Student ID	0,43	Moderate

Table 4.2: Inter-annotator agreement for individual attributes.

expected by chance. After inspecting the actual answers of the workers for these attributes, it shows that, for every instance of the attribute, one worker had a different answer. If we take a closer look at, for example, `Person`, the reason for disagreement could be that one of the workers views this attribute as privacy-sensitive, opposed to the other workers who view it as "safe", resulting in consistent disagreement. This also shows us that workers are consistent in their views about privacy and in applying it to images. If we look at both `Location` and `Date/Time`, which are both textual attributes, the same holds as well.

Taking all the above into consideration, we identified two main reasons why the crowd workers disagree with the ground-truth on certain classes. The first one is the co-occurrences of attributes. This means that certain attributes are often present in the image with other attributes and crowd workers deem one attribute "safe" when the accompanying attribute is removed, whereas when they are individually present in an image, they are always marked as privacy-sensitive. The other reason is that the crowd workers see certain attributes as "safe" in general, whereas the ground-truth says otherwise. In the subsequent sections, these two reasons are further looked into and explained.

Co-occurrence of Attributes

If we look at table 4.1, we see that some attributes are consistently ignored (or the other way around). Looking at 4.4, we see a few disabled people playing basketball. According to the dataset, there are three distinct privacy-sensitive attributes, namely: `Person`, `Physical Disability` and `Face`. If we look at how the crowd workers annotated this image, we see that the workers agree on the attribute `Face` as privacy-sensitive and identifying, but they disagree on the other two attributes, where only one worker selected `Person` and `Physical Disability`. However, in all the annotations of this im-

Figure 4.4: Example image in the dataset.

age, `Face` is marked as an (indirect) identifier, leading to a "safe" image when the `Face` elements are removed from the image.

Therefore, we look at the attributes that are consistently ignored and with which attributes they are present in an image. We will do this by looking at the annotations of images, where the attribute is present in the ground-truth of said image. This way, we know that the attribute must be present in the image, and see how the workers perceive it with surrounding attributes. We did this for the attributes in table 4.1 with a disagreement of +/- 40%, which are `Birth Date`, `Phone no.`, `Date/Time`, `Person`, `Handwriting` and `Passport`. For each attribute, we checked the image that they are present in and counted the different annotations that were made. In figure 4.5 you can see the attributes and the top-4 attributes that are annotated in that image.

What we can infer from this diagram is, take for example the attribute `Person`, when it is present in the image with attributes as `Face` and `Name`, the attribute itself is not deemed as privacy-sensitive. In this case, an attribute `Person` almost always is accompanied by the attribute `Face` and usually the face is seen as the identifying part of the body. This explains the large decrease in table 4.1. Another example is the attribute `Passport`, in the dataset there are two variants present, namely a passport with all the information visible, or the outside (cover) of a passport. In both cases, these get the privacy attribute `Passport`. However, for the first variant, the removal (or blurring, obfuscation, etc.) of the content of the passport might be enough for the

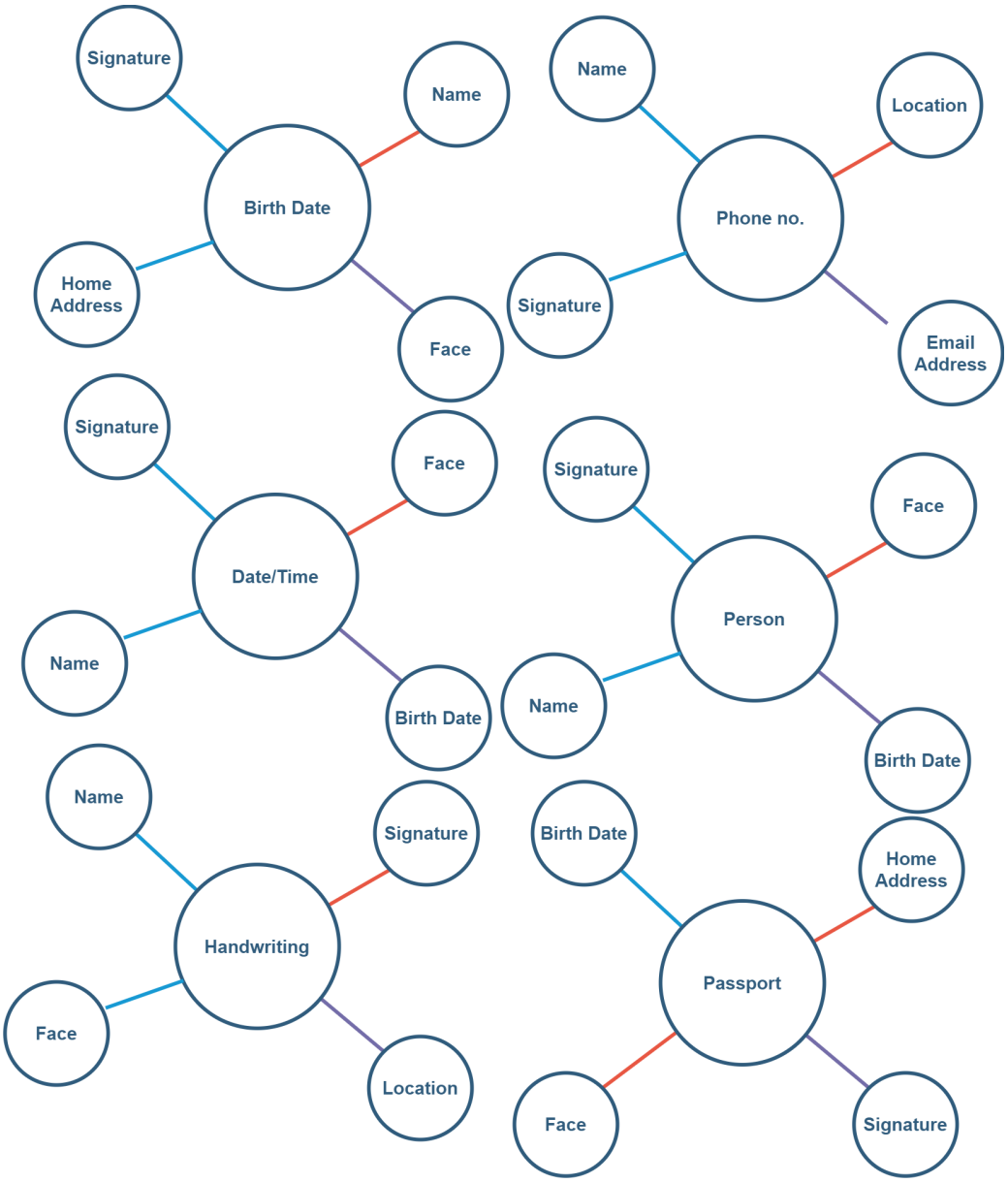


Figure 4.5: Co-occurrences of attributes

image to not be privacy-sensitive, as we can see from figure 4.5. As for the outside of a passport, the annotators consistently annotate it as not privacy-sensitive. The accompanying attributes are in these cases seen as privacy-sensitive, and even as (indirect) identifiers. Next to this, there are several attributes that dominate this diagram, which indicates that these are the attributes that are seen as most important, in the context of privacy-sensitive data.

Disagreement on Individual Attributes

Aside from the co-occurrence of attributes, there are also cases where the workers disagree on the importance of attributes, regardless of accompanying attributes. This means that the ground-truth has a different view on what is considered personal data than the crowd workers. From table 4.1, the comments that the workers placed and after inspecting the corresponding images, we see that several attributes have instances where the choice of the workers is justified. This applies to the attributes `Date/Time`, `Handwriting`, `Medical History` and `Passport`. For some instances of these attributes, it is unclear why they are marked as privacy-sensitive, as (solely from the image) it is not possible to identify a person.

Taking a closer look at one of the instances, which can be seen in figure 4.6, the ground-truth assigns two attributes to this image, namely `Passport` and `Location`. However, most of the workers did not select these attributes for the image, as only the cover of the passport is visible (with the country it is from). Therefore, the workers decide that it does not contain any privacy-sensitive information. If we compare this to other instances of the attribute `Passport`, we see that the inside of the passport, containing the full information of an individual, is visible. Thus, we can conclude that there is a discrepancy between the instances of this attribute, which leads to disagreement between the workers and the ground-truth. The same goes for the other attributes, if we look at `Medical History`, this attribute contains all the instances that has something to do with it. Looking at an instance of this attribute, see figure 4.7, the ground-truth labels this as `Medical History`, however, nothing important is visible in the image according to the crowd workers. Whereas, an instance containing a medical form does contain personal data, and just like for the attribute `Passport`, we identify discrepancies in the dataset.

4.3. Retrieving the Saliency Maps

In this section, we will look at the output of the machine learning model, and perform the crowdsourcing task, as described in sections 3.2.1 and 3.2.2. This section helps us with answering RQ3, which is: How do we characterize the strengths and weaknesses of a machine learning model for detecting personal data in images?

4.3.1. Assessing the Saliency Maps

The 100 images and their corresponding saliency for each attribute are inspected manually and the correctness of model task is done by us, as described in section 3.2.2. The amount of attributes present in these 100 images can be found in figure 4.9. As you

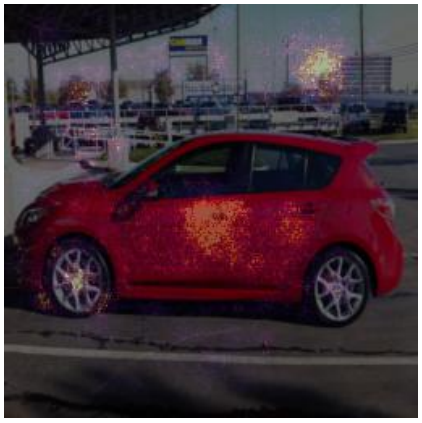
Figure 4.6: Instance of the attribute Passport in the dataset

can see, the attributes that have a high count, do also have a high count in the training dataset, which you can see in figure 4.1. Next to this, the model did not make predictions of certain attributes, which is mostly due to the low sample size in the dataset. Each of the saliency maps are rated on a scale from 1 to 5, where 1 means that the saliency map covers only a part of the attribute, 3 is a perfect coverage and 5 means that the attribute is covered, along with surrounding areas. Next to that, 0 means that the saliency map does not match the attribute at all. Alongside this rating, we also give the option to provide an explanation of the score given. In figure 4.8, four different instances of a face prediction are shown. Each of them got a different rating from the worker.

After the assessment was done, we looked at the scores of the different attributes. For this, we initially excluded the 0 scores(model either looks at completely wrong area, or attribute is not present in the image), as we first want to look at the attributes that are somewhat correctly predicted, i.e., looked at the right area by the model. This can be seen in figure 4.10. From the figure, we can see that most of the attributes have a score around 3, which means that the model looks at the correct areas when predicting that attribute. However, there are some attributes with a score lower than 2 and higher than 4. These are `Birth Date, Date/Time, License Plate, Location and Username`.

As for the predicted attributes that are predicted wrong, so a score given of 0, we look at how many instances of the attribute got this score and what the explanation was given by the workers. This can be seen in table 4.3. From the explanations, we can

Figure 4.7: Instance of the attribute Medical History in the dataset



(a) Rating 0, no face present at all.

(b) Rating 2, only a small part, particularly around the eyes, are highlighted.

(c) Rating 3, the saliency map covers the entire attribute.

(d) Rating 5, aside from the face, other parts are also highlighted.

Figure 4.8: Saliency maps of four face instances, each with a different rating provided by the worker.

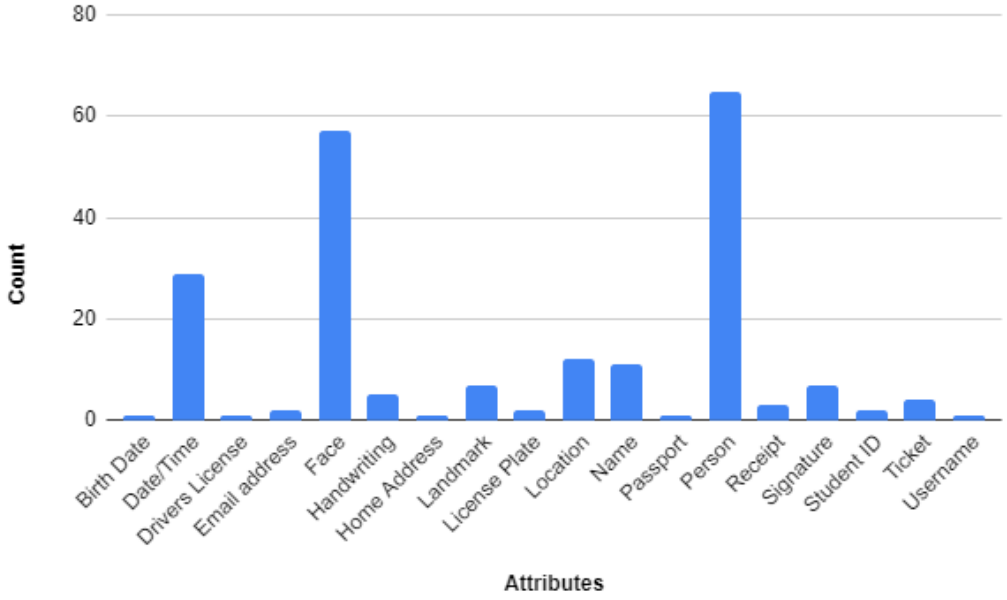


Figure 4.9: Attributes present in the result of the machine learning model.

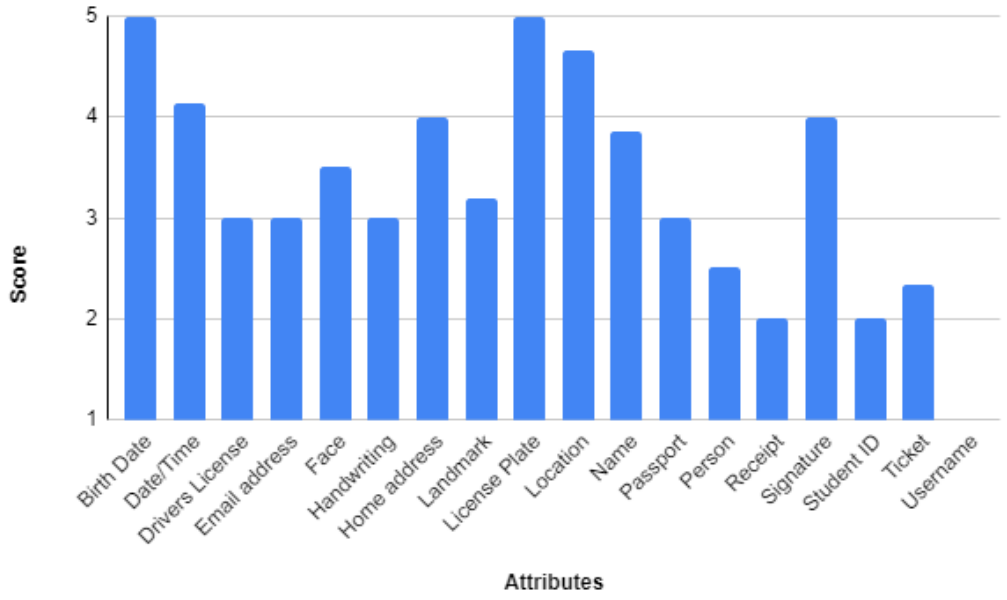


Figure 4.10: The scores of the different attributes, excluding 0 scores.

Attribute	Count	#Wrong Predictions	Explanation
Birth Date	1	0	
Date/Time	29	14	There are numbers present in the image.
Driver's License	1	0	
Email address	1	0	
Face	57	6	No face in image
Handwriting	5	0	
Home address	1	0	
Landmark	7	2	Might be because there is text present in the image
License Plate	2	1	No license plate, but there is a car visible in the image.
Location	12	9	Text present in image.
Name	11	4	Text present in image.
Passport	1	0	
Person	65	4	
Receipt	3	2	There is a ticket in the image, not a receipt.
Signature	7	4	Scribbles in images are recognized as signatures.
Student ID	2	1	
Ticket	4	1	Looks more like a business card in the image.
Username	1	1	

Table 4.3: Wrongly predicted labels and the reason given by the annotators.

deduce that textual attributes, such as Date/Time and Name, are difficult to recognize in images where other text appears. Whenever a number is present in the image, the model assigns the attribute Date/Time. The attributes handwriting and signature are often detected whenever there is a piece of handwritten text. The main flaw when detecting handwriting in an image that is a handwritten letter, is that the model only looks at part of the letter, even though the entire letter should be marked as handwriting.

What we can conclude from the assessment of the strengths and weaknesses of this machine learning model, is that the model is not fully able to distinguish different textual attributes. In the dataset, there are different textual attributes present and even combinations of textual attributes (Date/Time), which makes it difficult for the model to recognize patterns and therefore resulting in wrong classifications. As for the multimodal attributes, i.e., Student ID and Passport, there are not that many predictions made containing these attributes. Therefore, we can not say anything about the correctness of the predictions made, but we can say that the model has difficulties with assigning said attributes in the predictions. However, these attributes are frequently made up of different attributes, of which in turn we do have predictions. As for the visual attributes, the model seems to repeatedly predict them, although not fully correct all the time. Important to note is that the imbalance of the dataset also has a large impact on these predictions. The attributes that are often predicted, are also highly

present in the dataset.

5

Conclusion and Discussion

In this chapter, we will start with summarizing our results from the different experiments, discuss these results in order to answer the research questions. Furthermore, we will look at the limitations that we faced, discuss the implications our work has for future research, and what we want to do in the future.

5.1. Findings

In this section, we will summarize our findings, starting with the results about the notion of privacy, where we try to conceptualize privacy-sensitive elements in the context with the help of humans (RQ2). Next, we discuss the results of the saliency map assessments, where we try to characterize the strengths and weaknesses of a machine learning model for detecting privacy-sensitive elements in images (RQ3). Lastly, we combine the findings in order to find out to what extent can we use humans efficiently to increase the detection of privacy-sensitive elements in images? (RQ)

5.1.1. Notion of Privacy

From the notion of privacy task, there are several findings that we can discuss. The first one is that we see that the agreement of workers, on what is privacy-sensitive in images, does not exceed a moderate strength (see table 4.2), but it is also interesting to see that workers individually are consistent in applying their views about privacy, similarly to the work presented in Han [17], and therefore consistently disagreeing with other workers that have different views about privacy. Comparing the moderate agreement between workers with Greenberg [16], where they say that legal interpreters have trouble with reaching a consensus regarding on how to interpret law, the same holds for "ordinary" people.

Next to this, we see that there are mainly two reasons why workers differentiate from a ground-truth, which are the co-occurrence of attributes and a plain disagreement in view of what is considered privacy-sensitive. Certain attributes hold more value than others, some attributes might individually not reveal any information about

a person, but together with other attributes, they might. Therefore, it is important to not only look at individual attributes present in images, but look for certain combinations of attributes and maybe adding some weights to attributes in order to differentiate between the importance of certain attributes. This might be useful in applications that give a score, on how privacy-sensitive an image is.

From the results, we also see that there is often a disagreement on an individual attribute between the ground-truth and the workers. This stems from the fact that an attribute, take passport for example, has many instances, and not all of them are privacy-sensitive (see figure 2.1, even though the attribute is correct. Therefore, we can conclude that this one-fits-all approach is therefore not suitable, and calls for a better distinction within attributes of instances that can be considered safe and ones that are not, adding another dimension to the dataset.

5.1.2. Saliency Maps

Looking at the results of the assessment of the saliency maps, we see that the model does not distinguish the textual attributes easily. If we take a closer look at one attribute that has many samples and still underperforms, namely `Date/Time`, we see that this is a combination of two different attributes. Next to this, date and time have many ways of writing it down, so it would benefit from at least separating them. The other attributes, besides textual attributes, that have a lot of samples in the dataset, seem to be predicted fine, even with a relatively general model, and the coverage by the saliency maps also reflects that. For the other attributes, the model clearly struggles with them, which is largely due to the fact that there are not that many samples present in the dataset.

5.1.3. Synthesis

Therefore, we can say that humans can be efficiently used to increase the detection of privacy-sensitive elements in images. Looking at the notion of privacy, we see that workers have different views on what they consider privacy-sensitive, and that they are consistent in applying their views on different images. The disagreement can be attributed to several factors, which will be further discussed in the next section, but overall, this provides great insights on how people apply privacy laws to images.

For the saliency maps, we see that humans are capable of assessing the workings of a model, providing good results and detailed feedback, whilst also increasing trust in the model, similarly to [26] and [30]. We also see that the added context of privacy is not a restricting factor, and the workers are capable of performing the task within this context.

The performance of the workers shows us that we do not need solely experts in the field for applying privacy laws to images. This means that future researchers can save costs if they choose crowd workers over experts (or a combination of both) within this context, while keeping the performance loss at a minimum.

5.2. Limitations

There are several limitations that we want to address, starting off with the imbalance of the dataset. As noticed in the previous section, this has a large impact on how well the model performs, and therefore not producing the optimal intermediates for us to properly assess the strengths and weaknesses of the machine learning model. In table 4.3 we see this imbalance back in the results, numerous attributes have a few results, thus it is not possible to draw conclusions from these annotations for these specific attributes.

Next to this, there is a large diversity of attributes, which also leads to a low number of samples for some attribute, narrowing down the scope to several attributes would probably lead to a better understanding of the machine learning model. On the other hand, there is also a need for a better distinction within attributes, increasing the number of attributes. Therefore, we need to find an optimal selection of attributes that are distinct enough, while keeping the number of attributes at a minimum.

Furthermore, there are also a lot of attributes which have instances that are not necessarily privacy-sensitive, this leads to incorrect classifications, which decreases the trust a user has in the model, and possibly leads to not using the model at all [13]. An example can be seen in figure 4.6, where the instance is labelled as `Passport`, and therefore considered privacy-sensitive. However, most workers did not see this particular instance as privacy-sensitive. Other instances of `Passport` show the inside of it, with the full details of an individual, is indeed considered privacy-sensitive by the workers. This calls for a better distinction within attributes itself, perhaps by introducing a new dimension that says whether the attribute is safe or not.

5.3. Implication

From the limitations, we see that there is a need for a dataset, that needs to be balanced in terms of attributes, separates several attributes and makes distinctions within attributes. This is not a simple task, and needs to be carefully done, in order to create a high-quality dataset, which will benefit future researchers.

Next, what we see is that there are links between several attributes, combinations of them can be used, in the context of detecting privacy-sensitive elements in images, to identify a person, whereas they individually might not identify a person. This paves the way for a more sophisticated approach to detect privacy-sensitive elements in images, opposed to looking at attributes individually.

Regarding the notion of privacy, we see that, regardless of the quality of the dataset, people do not necessarily hold the same view on what is privacy-sensitive or not, which is in line with Greenberg [16]. However, they are consistent in their judgement, meaning that they are capable of interpreting a privacy definition and consistently applying it to, in this case, images.

5.4. Future Work

As for our own future research, there are several things that we might dive into. The first thing is to conduct a larger crowdsourcing task, for both the notion of privacy and assessing the saliency maps, with different demographics included. For the notion of privacy, the inclusion of people around the world may provide us with insights about how people's background and country of origin impacts the interpretation of a privacy definition, which in turn can be useful for creating different rule sets for different parts of the world. A larger number of crowd workers, will possibly result in a more solid foundation of the results that we derive.

Next to this, it would also be interesting to see whether this approach still holds with different settings, for instance using other datasets and other models, to see whether the approach generalizes well. With these different settings, we can also compare them with each other and see which setting has the best results.

Lastly, in the crowdsourcing task about the notion of privacy, we only present the workers images that are publicly available. An extension of this task, where people also have to judge whether a privacy-sensitive element is present in their own uploads, similarly to Han et al. [17], would show us whether the workers are still consistent in their judgements when they see an example of their own profile. Essentially, it would be interesting to see whether the privacy paradox [28], holds in this scenario.

Bibliography

[1] Eu data protection directive 95/46/ec. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046.

[2] Interpretation. URL https://legal-dictionary.thefreedictionary.com/Interpretation.

[3] Privacy act of 1974, Apr 2021. URL https://www.justice.gov/opcl/privacy-act-1974.

[4] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[5] Abdullah Alshaibani, Sylvia Carrell, Li-Hsin Tseng, Jungmin Shin, and Alexander Quinn. Privacy-preserving face redaction using crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):13–22, Oct. 2020. URL https://ojs.aaai.org/index.php/HCOMP/article/view/7459.

[6] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. What do you mean? interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*, pages 1937–1948, 2021.

[7] Aharon Barak. *Purposive Interpretation in Law*:. Princeton University Press, 2011. ISBN 978-1-4008-4126-4. doi: 10.1515/9781400841264. URL https://doi.org/10.1515/9781400841264.

[8] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[9] Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian*, 17:22, 2018.

[10] Sophie Cockcroft and Saphira Rekker. The relationship between culture and information privacy policy. *Electronic Markets*, 26, 07 2015. doi: 10.1007/s12525-015-0195-9.

[11] Anubrata Das, Brandon Dang, and Matthew Lease. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 33–42, 2020.

[12] Knut De Swert. Calculating inter-coder reliability in media content analysis using krippendorff's alpha. *Center for Politics and Communication*, 15, 2012.

[13] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[14] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. 09 2013. doi: 10.1145/2499149.2499168.

[15] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[16] Mark Greenberg. Legal Interpretation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[17] Kyungsik Han, Hyunggu Jung, Jin Yea Jang, and Dongwon Lee. Understanding users' privacy attitudes through subjective and objective assessments: An instagram case study. *Computer*, 51(6):18–28, 2018.

[18] Jeff Howe et al. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[19] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey Bigham, Jaime Teevan, Ece Kamar, and Walter Lasecki. Crowdmask: Using crowds to preserve privacy in crowd-powered systems via progressive filtering. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 5(1):89–98, Sep. 2017. URL `https://ojs.aaai.org/index.php/HCOMP/article/view/13314`.

[20] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[21] Greg Little and Yu-An Sun. Human ocr: Insights from a complex human computation process. In *Workshop on Crowdsourcing and Human Computation, Services, Studies and Platforms, ACM CHI*. Citeseer, 2011.

[22] Erika McCallister. *Guide to protecting the confidentiality of personally identifiable information*, volume 800. Diane Publishing, 2010.

[23] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images, 2017.

[24] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[25] Harshvardhan Pandit, Axel Polleres, Bert Bos, Rob Brennan, Bud Bruegger, Fajar Ekaputra, Javier Fernández, Roghaiyeh Hamed, Elmar Kiesling, Mark Lizar, Eva Schlehahn, Simon Steyskal, and Rigo Wenning. *Creating a Vocabulary for Data Privacy: The First-Year Report of Data Privacy Vocabularies and Controls Community Group (DPVCG)*, pages 714–730. 10 2019. ISBN 978-3-030-33245-7. doi: 10.1007/978-3-030-33246-4_44.

[26] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL `http://arxiv.org/abs/1602.04938`.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[28] Monika Taddicken. The 'privacy paradox'in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure. *Journal of Computer-Mediated Communication*, 19 (2):248–273, 2014.

[29] Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1):7026–7071, 2017.

[30] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis, 2017.